Integrating Materials and Manufacturing Innovation
a SpringerOpen Journal

REVIEW Open Access

Making materials science and engineering data more valuable research products

Charles H Ward[1*], James A Warren[2] and Robert J Hanisch[2,3]

* Correspondence: charles.ward.4@us.af.mil
[1]Air Force Research Laboratory, Materials and Manufacturing Directorate, Wright-Patterson AFB, OH 45433, USA
Full list of author information is available at the end of the article

Abstract

Both the global research community and federal governments are embracing a move toward more open sharing of the products of research. Historically, the primary product of research has been peer-reviewed journal articles and published technical reports. However, advances in information technology, new 'open access' business models, and government policies are working to make publications and supporting materials much more accessible to the general public. These same drivers are blurring the distinction between the data generated through the course of research and the associated publications. These developments have the potential to significantly enhance the value of both publications and supporting digital research data, turning them into valuable assets that can be shared and reused by other researchers. The confluence of these shifts in the research landscape leads one to the conclusion that technical publications and their supporting research data must be bound together in a rational fashion. However, bringing these two research products together will require the establishment of new policies and a supporting data infrastructure that have essentially no precedent in the materials community, and indeed, are stressing many other fields of research. This document raises the key issues that must be addressed in developing these policies and infrastructure and suggests a path forward in creating the solutions.

Keywords: Materials data; Data policy; Data repository; ICME; MGI; Integrated Computational Materials Engineering; Materials Genome Initiative; Data archiving

Introduction

Reliance on shared digital data in scientific and engineering pursuits - whether the data are derived from computation or experiment - is becoming more commonplace within the materials science and engineering (MSE) community. Concurrently, government policies across the globe are embracing an 'open science' model which sets a requirement for sharing digital data generated from publicly funded research. A recent joint Materials Research Society and The Minerals, Metals and Materials Society (MRS-TMS) survey on 'big data' in materials science and engineering showed that 74% of respondents would be willing to participate in sharing their data if it was encouraged as a term and condition of funding or publishing, assuming the proper safeguards were in place [1]. However, it is fair to say that the MSE community currently lacks the strategy, framework, standards, and culture needed to support materials data curation and sharing. A unified approach is needed to meet the growing demands of the community and a plan to meet government requirements for broad access to digital data. It is clear that the

Springer

United States government work, not subject to copyright.; licensee Springer. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited.

peer-reviewed journals and government-sponsored technical reports serving the MSE community can be an essential component to the solution, and there is now an opportunity to proactively plan how they may best serve the growing needs of their constituency.

We have structured this paper to, first, provide the reader a general awareness of the global environment and ongoing activities concerning the management of research data. We then present a perspective on the benefits of data archiving to the MSE community and outline what attributes and characteristics a data archiving solution should have. We follow this with a discussion of key challenges yet facing the establishment of a digital materials data infrastructure. Finally, we propose a way forward to tackle the creation of community-based solutions for data archiving policies and data repositories.

Review

Global context

The 2008 NRC report on Integrated Computational Materials Engineering (ICME) highlighted the importance digital data will play in the future of materials science and engineering [2]. MSE's ever increasing reliance on computational modeling and simulation will demand digital data as the feedstock for solutions in both science and engineering.

In the USA, the National Institutes of Health have long promoted a policy of open access to data generated from their grants [3]. In the mid-1990s, the Human Genome Initiative spawned the Bermuda Principles which called for immediate public posting of sequences of the human genome [4]. More recently, the National Science Foundation has adopted a requirement that applicants provide a data management plan in grant proposals [5]. Specific to the materials community, the sharing of digital data is a key strategy component of the US's Materials Genome Initiative (MGI), and mechanisms to foster and enable sharing are actively under consideration [6].

The European Union has been very proactive in studying the impacts of a digitally linked world on the scientific community. The EU Framework Programme 7 funded a project called Opportunities for Data Exchange that has produced several relevant reports on publishing digital data in the scientific community [7]. And, in June 2012, the Royal Society published 'Science as an open enterprise' which promotes free and open access to scientific results, including data [8]. These studies are now broadly informing government policy. For example, recent policy issued in the UK in July 2012 calls for government-funded research to be published in open access journals and requires access to supporting research data [9]. In February 2013, Dr. John Holdren, director of the Office of Science and Technology Policy (OSTP), issued a directive to all federal agencies to develop plans to make the results of federally funded research more accessible to the public [10]. A key component of this directive is a call for agency plans to include a means by which digital data resulting from research can be made available to the public. In support of this policy, the White House has established a useful web site providing resources supporting the establishment of open data [11]. The US Government funding agencies have since provided their plans to address OSTP's open research policy, and results are imminent.

Other technical communities have addressed the challenges of access to digital data with a variety of approaches. Indeed, the biology community has implemented a number of differing approaches; for example, the approach taken in genetics versus that

adopted by evolutionary biology [12,13]. In other disciplines, one subfield of thermodynamics has already adopted a very structured approach to archiving data, while the earth sciences community has embarked on an effort to define its approach [14,15]. The astronomy community has dedicated international resources to the development of the Virtual Observatory, an infrastructure that enables global data discovery and access across hundreds of distributed archives [16]. Despite the differing mechanics of implementation, all the approaches were rooted in a community-led effort to define the path best suited for that particular technical field.

In response to these trends, technical communities and publishers have developed and implemented open access journals and data archiving policies. Again, the field of biology appears to be leading the way on both these fronts. One example of this trend is *Database: The Journal of Biological Databases and Curation*, an open access journal dedicated to the discussion of digital data in biology [17]. And in a recent development, Nature Publishing Group has launched a new open access journal, titled *Scientific Data*, which is dedicated to publishing descriptions of scientific datasets and their acquisition [18]. It will initially focus on life, biomedical, and environmental science communities. Furthermore, the Public Library of Science recently strengthened its policy on data access: 'PLoS journals require authors to make all data underlying the findings described in their manuscript fully available without restriction, with rare exception' [19].

In order to begin a dialog within the MSE community, the National Institute of Standards and Technology (NIST) convened a workshop on digital materials data in May of 2012 under the auspices of MGI. The workshop identified a number of barriers that need to be addressed during creation of a data strategy for materials, they include: materials schema/ontology, data and metadata standards, data repositories/archive, data quality, incentives for data sharing, intellectual property, and tools for finding data [20]. Other disciplines, notably evolutionary biology, have demonstrated that peer-reviewed journals have the potential to contribute solutions to these barriers to data sharing [21].

Benefits of archiving materials science and engineering data

There is a growing realization within the global scientific community that the data generated in the course of research is an oft overlooked asset with considerable residual value to other scientists and engineers and that often a significant portion of the data is stored but not accessible. The following are several anticipated benefits of increasing access to materials science and engineering data in digital form:

Data reuse

- Scientific productivity and return on investment in research infrastructure.
- Secondary hypothesis testing.
- Reducing/eliminating paying for data generation multiple times.
- Comparing with previous studies.
- Integrating with previous and future work.
- Reproducing and checking analyses.
- Simplifying and enhancing subsequent systematic reviews and meta-analyses.

- Facilitating interdisciplinary research.
- Teaching.

Incentives

- Increasing academic credit (citations).
- Access to one's own data at a future date.
- Convenience and security of cloud storage.

Others

- Testing algorithms/computations with validated reference datasets.
- Meeting funding agency requirements to share data.
- Reducing the potential for duplication of effort.
- Reducing of error and fraud.

The MRS-TMS 'big data' survey asked participants to evaluate whether given attributes would act as impediments or motivators to sharing data, Figure 1 [1]. The bottom of the graph shows that the largest impediments are primarily driven by legal considerations. The top of the graph demonstrates that the strongest positive motivators are the increased attention and credit a researcher may draw for one's work. It is clear that widespread sharing of digital materials data will require not only technological advances but also cultural shifts that include modernization of traditional incentives for the sharing of scholarly works to include recognition for publication of data. While there is no universally accepted solution available at present, new tools such as the Thomson Reuters Data Citation Index may help provide avenues for this recognition [22].

The impact on research productivity owing to the provision of well-calibrated, well-documented, archival data products is clearly demonstrated in the case of NASA's Hubble Space Telescope (HST). Initially, archival data was not used very extensively; the data suffered from spherical aberration, of course, resulting in a factor of approximately ten decrease in sensitivity from expectations. But in the early 1990s, there was

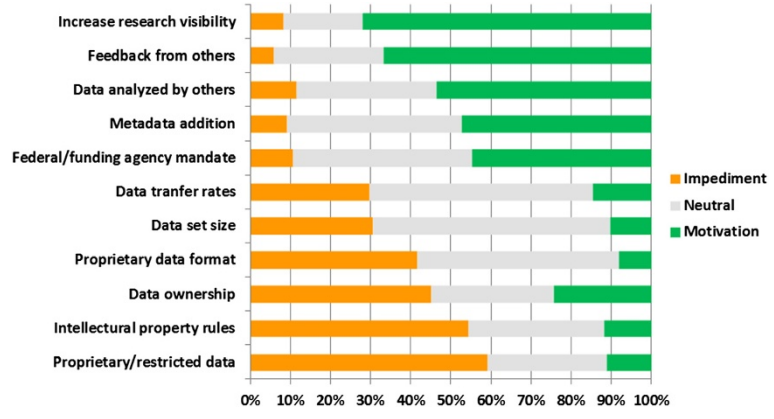

Figure 1 MRS-TMS big data survey result. Responses to the question: 'Do you consider the following items to be impediments or motivation for you to share your data with the world?' The abscissa depicts the response rate to the three choices: impediment, neutral, or motivation.

also somewhat of a stigma attached to using archival data for research: this was somehow not as good or pure as collecting one's own data at a telescope. But times have changed, and HST archival data is now used extensively by astronomers unaffiliated with the teams making the original observations. The resulting research papers account for more than half of all peer-reviewed publications based on HST observations (see Figure 2). There are a number of reasons for the big increase in archival data use. HST observing time is very difficult to get, with typically a seven-to-one oversubscription ratio in the proposal process. All HST data is routinely pipeline processed, yielding an archive of 'science ready' data products. All HST data becomes public after a nominal 12-month embargo period. And HST data taken for one purpose can often be utilized for studies of a substantially different intent. While this high level of reuse may not be achieved for all research experiments, the HST example clearly shows that a substantial improvement in research productivity can be achieved, at a very modest incremental cost, when proper care is taken in designing the data management system.

In materials science, there is a strong case that data obtained at great public expense should be made available to as large a group of researchers as is practical, noting that, unlike astronomical data, there can be important constraints due to both national security and intellectual property concerns. Acknowledging these issues, however, materials data obtained from national user facilities, such as data obtained by scattering of synchrotron light or neutron beams, are examples of a scarce yet valuable information stream. Indeed, these facilities have recognized the importance of good archiving capability but have not focused on distribution of these data. Making such information more widely available would increase the amount of materials knowledge discovery with a small investment relative to the costs of repeating the work.

Background for data archiving

A common approach to archiving materials data has several benefits, but its primary value would be to provide unified, consistent guidance and expectations throughout the scientific and engineering community. However, while the development of an archiving policy itself may be relatively straightforward, the infrastructural issues

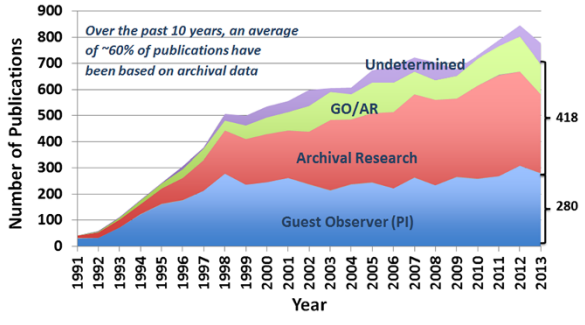

Figure 2 Use and reuse of archived data from the Hubble Space Telescope. HST data are used approximately twice as often in research papers written by scientists with no connection to the original investigators proposing the research. This more than doubles the productivity of HST at a marginal extra cost of providing well-calibrated data in an easy to access archive.

necessary to support policy implementation are extraordinarily complex. These issues include the establishment of viable

- Repositories for materials data.
- Standards for data exchange.
- Citation and attribution protocols.
- Data quality metrics.
- Intellectual property and liability determinations.

Characteristics of an archiving solution

In order for a data archiving solution to be of lasting value to researchers and maintain the rigorous, archival standards of relevant publications, it should have the following minimum set of characteristics:

- Persistent citation.
- Data discoverability.
- Open access (for journals).
- Ease of use.
- Minimal cost.

To archive or not to archive?

The most critical question to be answered in setting policies for publications is 'what data should be archived?' The answer is essential in providing clear expectations for authors, editors, and reviewers, as well as determining the size of the data repositories needed. Other disciplines have already embarked on this journey and have devised a variety of approaches that suit the data needs of their communities for their stage of 'digital maturity.' Two ends of the spectrum in addressing this question are presented here. The first assumes all data supporting a publication are worthy of archiving. This criterion is found most often in peer-reviewed journals that have narrow technical scope and generally deal with very limited data types. For example, journals in crystallography and fluid thermodynamics have very stringent data archiving policies that prescribe formats and specific repositories for the data submitted [23,24]. Other journals that cover broader technical scope, and therefore deal with more heterogeneous data, have implemented more subjective criteria for data archiving and a distributed repository philosophy. Earth sciences and evolutionary biology have typically taken this approach. It is likely that the approach adopted by MSE publications may also span a similar spectrum, depending on the scope of the publication.

The MRS-TMS 'big data' survey provided insight into the community's perspective on the relative value of access to various types of materials data, shown in Figure 3. It is interesting to note that as the complexity of the data and metadata increase (generally) toward the right-hand side of the chart, the community's perceived need to have access to this data decreases. This could be due to many factors including the difficulty in assuring the quality of such data as well as the lack of familiarity with tools to handle the data complexity. However, with complexity comes a richness of information that if properly tapped could be extraordinarily valuable. In astronomy, for example, the Sloan Digital Sky Survey created a very complex database of attributes of stars, galaxies, and

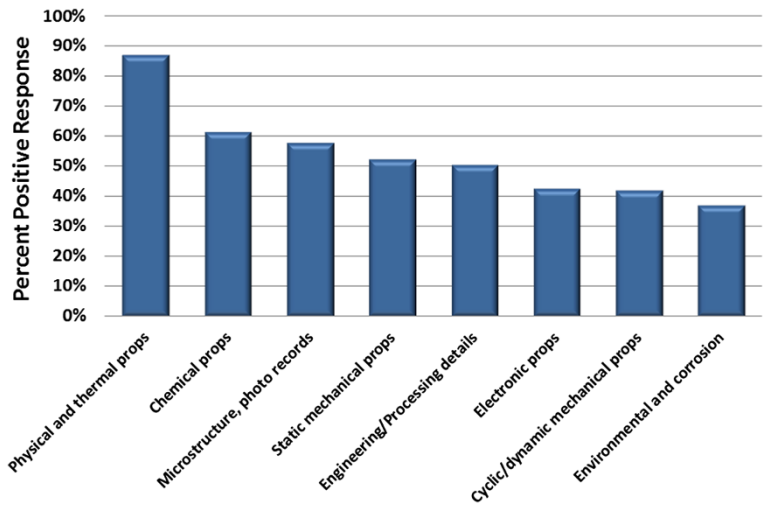

Figure 3 MRS-TMS big data survey result. Responses to the question: 'What scientific/technical databases and data mining tools would be most useful if they could be created?'.

quasars. The wealth of information and immense discovery potential led many in the research community to become expert users of SQL and for the survey to yield nearly 6,000 peer-reviewed publications.[a]

For those publications with wide technical scope, it will be difficult to provide a universal answer to 'what data should be archived?' In these cases, the decision for what data to archive may best be left to the judgment of the authors, peer reviewers, and editors. A particularly useful metric might be the cost/effort to produce the data. For example, the 'exquisite' experimental data associated with a high-energy diffraction microscopy experiment provide very unique, expensive, and rich datasets with great potential use to other researchers [25]. Clearly, based on these factors, the dataset should be archived. On the other hand, the results from a model run on commercial software that takes 5 min of desktop computation time may not be worthy of archiving as long as the input data, boundary conditions, and software version were well defined in the manuscript. Of course, one must account for the perishable nature of code, particularly old versions of commercial code. However, even the data from the common tensile test may be worthy of archiving as publications do not typically provide the entire curve; while the paper may report only yield strength, another researcher may be interested in work hardening behavior. Having the complete dataset in hand allows another researcher to explore alternative facets of the material's behavior. The basic elements of the criteria for determining the data required for archiving could include the following:

- Are the data central to the main scientific conclusions of the paper?
- Are the data likely to be usable by other scientists working in the field?
- Are the data described with sufficient pedigree and provenance that other scientists can reuse them in their proper context?
- Is the cost of reproducing the dataset substantially larger than the cost of archiving the fully curated dataset?
- Is the dataset reproducible at all, or does it stem from a unique event or experiment?

Data itself can come in a variety of 'processed' levels including 'raw', 'cleaned', and 'analyzed'. Such characterizations are subjective, though some disciplines have adopted quite rigorous definitions. Nonetheless, given the diversity of materials data, care will need to be taken in determining the appropriate amount of processing performed on a dataset to be archived. While raw or cleaned data is much preferred for its relative simplicity in reuse, it is probably much more important at this stage of our digital maturity that the metadata accompanying the dataset provide sufficient pedigree and provenance to make the data useful to others, including definition of the post-acquisition (experiment or computation) processing performed.

Another factor to consider in setting the guidelines for which data need to be archived is the expected annual and continuing storage capacity required. A very informal survey of 15 peer-reviewed journal article authors in NIST and Air Force Research Laboratory (AFRL) found that most articles in the survey had less than 2 GB of supporting data per paper. Currently, the time and resources required to upload (by authors) and download (by users) data files less than 2 GB are quite reasonable. However, those papers reporting on emerging characterization techniques such as 3D serial sectioning and high-energy diffraction microscopy were dependent on considerably larger datasets, approximately 500 GB per paper. Other disciplines have established data repositories to support their technical journals. Experience to date indicates that datasets of up to approximately 10 GB can be efficiently and cost effectively curated. Repositories such as the Dryad digital repository show that datasets of this magnitude can be indefinitely stored at a cost of $80 or less [13]. However, datasets approaching 500 GB will very likely require a different approach for storage and access. Thus, a data repository strategy needs to consider this range in distribution of datasets. An additional factor when considering long-term storage requirement is the high global rate of growth in materials science and engineering publications. Figure 4 shows the dramatic growth in the number of MSE journal articles published over the past two decades, indicating a commensurate amount of accompanying data.

Data repositories

Aside from crystallographic data repositories, there are at this time perhaps no dedicated materials data repositories that meet the required characteristics defined above. The materials science and engineering community does have numerous publicly accessible data repositories; however, the majority of these are associated with specific projects or research groups, and their persistence is therefore dependent on individual funding decisions. These repositories are primarily established to house and share the research data generated within a specific project or program. They generally do not follow uniform standards for data and metadata nor provide for data discoverability and citation. There are very few repositories established with the explicit objective of providing MSE with public repositories for accessible digital data. In short, publicly accessible, built-for-purpose repositories and the associated infrastructure for access, safe storage, and management still need to be developed and sustainably funded - this is the largest impediment to implementing viable data archiving policies. (See, for example, 'sustaining domain repositories for digital data: a white paper' [26]). In establishing their Joint Data Archiving Policy, journals in evolutionary biology did not prescribe specific repositories; instead, they allowed a mix of repositories to be used by authors as long as

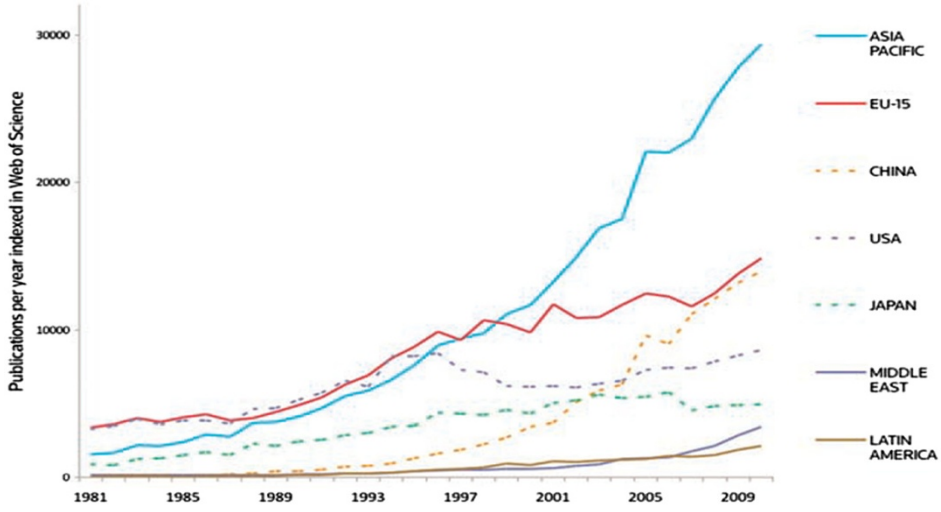

Figure 4 Number of materials science and engineering publications per year. Source: Thomson Reuters Web of Knowledge[SM].

they met established criteria. Such criteria may be as simple as requiring data cited to be permanently archived in data repositories that meet the following conditions:

- Publicly accessible throughout the world.
- Committed to archiving data sets indefinitely.
- Allow bi-directional linking between paper and dataset.
- Provide persistent digital identifier.

One tempting option might be to take advantage of the online storage capability several journals already offer for supplementary materials accompanying journal articles. However, as presently constructed, these are not amenable to best practices for dataset storage as they generally are not independently discoverable, searchable, separately citable, nor aggregated in one location. In fact, some publishers are reducing or eliminating supplementary file storage due to the haphazard structure and rules associated with their use. Further, new global government policies promoting open access to research works have the publishing industry in a state of flux with regard to their long-standing, subscription-based business model. Publishers have been reticent in taking on a data archiving responsibility given the economic uncertainties in the publishing marketplace.[b] Also, there is a possibility that some for-profit publishers could try to restrict access to digital data assets that are co-located with the journal.

As alluded to in the previous section, a fundamental consideration in repository design and/or selection is the level to which the repository will present structured versus unstructured data. Structured technical databases tend to be more useful to a technical community due to their uniformity, as evidenced by their data reuse rate.[c] A perfect construct would see the vast majority of materials data resident within structured repositories. A disciplined data structure provides enormous advantages to the researcher both in terms of data discoverability and confidence in its use. However, this structure must be enabled by the application of broader and deeper standards for data and metadata, standards that do not currently exist.

In all likelihood, as in biology, MSE publications will be dependent on a collection of repositories that are tailored to specific materials data. For example, NIST is building and demonstrating a data file repository for CALPHAD (Calculation of Phase Diagrams) and interatomic potentials [27]. These may be expandable and largely sufficient for thematic publications such as those devoted to thermodynamics and diffusion. However, repositories such as this will only fill a relatively small niche need in MSE. *Integrating Materials and Manufacturing Innovation* is piloting an effort to link articles with their supporting data using the NIST repository according to the criteria outlined above, an example can be found in an article by Shade et al. [28,29].

Finally, a business model for sustainably archiving materials data is required. Other technical fields, such as earth sciences, can at least partially rely on government-provided repositories for large and complex datasets. Without these types of repositories to build on, MSE will need to establish viable repository solutions. In response to funding-agency requirements for data management plans, some universities, Johns Hopkins for example, are beginning to provide centrally hosted data repositories, but these are not yet common [30]. Private fee-for-service repository services, such as LabArchives and Figshare, are also evolving to meet growing demand for accessible data storage [31,32][d]. Additionally, ASM International is working to create a prototype materials data repository through its close association with Granta Design. Termed the Computational Materials Data Network (CMDN), this is a promising option as the data repository will provide a structured database specifically for materials data; but the business model for CMDN has not yet been solidified [33]. A key open question remains how funding agencies will respond to the OSTP open research policy memo, and how they will fund activities making data open to the public.

Standards enabling data discoverability, exchange, and reuse

As noted in the previous section, standards for data and metadata provide the basis for a structured data archive, enabling the rapid discovery of data and assisting in determining the data's relevance and usefulness. At the most basic level, good data practice generally requires the generation, and acceptance, of a vocabulary defining the terms used to describe reported data. This assures the data user that they precisely understand the context of the data they are reviewing. From this level, other attributes, features, or requirements can be levied on a data management system including ontologies, schema, and formats [34].

Other fields have studied these issues as a community, and MSE is now reinvigorating a concerted effort to define its approach to setting data standards. Serious efforts to address standards for materials data, particularly structural materials data, were undertaken as long as 30 years ago [35]. In 1985, ASTM International established Committee E49 on Computerization of Material Property Data to develop standard guidelines and practices for materials databases [36]. ASTM International devoted quite substantial effort over a decade and issued data standards relevant to materials through the 1990s, specifically addressing key issues such as how to describe materials, how to record data, data quality indicators, harmonization of terminology, and guidelines for building and distributing databases. The standards have been since withdrawn, but those such as ASTM E1314-89, 'Practice for Structuring Terminological Records Relating to Computerized Test Reporting and Materials Design Formats', are clearly in need in today's environment - though in more web-enabled format [37]. ASTM International has been reviving its efforts to

provide guidance on the digitization of materials test data by exploring the re-establishment of its computerization and networking of materials databases symposium series [38]. The European Union is studying the creation of standards for the exchange of engineering materials data through the European Committee for Standardization [39]. The target for these standards is structural materials with an early emphasis on aerospace applications. And the European Commission is funding a broader activity called the *Integrated Computational Materials Engineering expert group* (ICMEg) with the aim of developing the standards and protocols needed to support the digital exchange of materials data needed to conduct ICME [40]. Several recent papers have proposed standards for other types of materials data to include thermodynamic and image-based data [41,42]. There are also closed-loop approaches to materials data standardization that exist within commercial data management software packages, but these are not generally available to the public.

While the field of information technology is continuously evolving to provide solutions to more productively using unstructured data, at present there is no community-wide accepted practice for MSE data and metadata standards. Near-term solutions for governing the archiving of materials data will need to be relatively loose, flexible, and evolutionary with a drive toward more standardization. While publishers may not be able to directly provide data repository services, they are well positioned and willing to aid the community in establishment of data standards. In the pursuit of standardization across a technical field, Michael Whitlock, a primary champion of journal data archiving in the field of evolutionary biology, offered this quote from Voltaire based on his experience: 'the perfect is the enemy of the good'.[e] It is perhaps much more important at this stage of our digital maturity that MSE first implement data archiving with the best guidance available and work to build in standardization over time.

Data citation and attribution

Well-developed and uniform data citation standards are required to ensure that linkages between publications and datasets are enduring and that creators of digital datasets receive appropriate credit when their data are used by others. Standards for data citation practices and implementation provide the mechanism by which digital datasets can be reliably discovered and retrieved. Closely related to data citation, other challenges include the ability to reliably identify, locate, access, interpret, and verify the version, integrity, and provenance of digital datasets [43]. Any data archiving policy must concern itself not only with how publications should appropriately cite the datasets used but must also require attribution to authors of datasets outside the document.

Numerous organizations in the EU and USA have studied this issue and are continuing to refine technology solutions and best practices. For example, CODATA and the National Academy of Sciences released an in-depth international study and recommendations on citation of technical data [44]. Recently, these transnational initiatives have coalesced to produce a unified Joint Declaration of Data Citation Principles that is appropriate for any type of technical publication [45]. The eight principles define the purpose, function, and attributes of data citations and address the need for citations to be both understood by humans and processed by machines. With a slightly different perspective focused more on the mechanics of linking published articles with data repositories, DataCite and the International Association of Scientific, Technical and Medical

Publishers have issued a joint statement recommending best practices for citation of technical datasets in journals [46]. Two of the key recommendations include encouraging authors of research papers to deposit researcher validated data in trustworthy and reliable data archives and encouraging data archives to enable bi-directional linking between datasets and publications by using established and community-endorsed unique persistent identifiers such as database accession codes and digital object identifiers (DOIs).

An outstanding technical issue in data citation yet to be resolved concerns the granularity of the datasets used in a publication, both spatially and temporally. Spatial granularity refers to a subset of the dataset used in the research. Temporal granularity can refer to either the version of the dataset used or the temporal state of the dataset used if the dataset itself is dynamic.

Data quality

A key concern in linking datasets to publications is the provision of quality metrics; that is, can the data's ultimate reliability be assessed in a meaningful manner? Materials data can be provided as two basic types: experimental and computational; both types assume underlying models. In order for data and these associated models to be usable, their quality must be ascertained. In this context, it is useful to define the following for data and models:

- Pedigree - where did the information come from?
- Provenance - how was the information generated (protocols and equipment)? This metadata should be sufficient to reproduce the provided data.

In addition to these qualitative descriptors of the data, there are a number of meaningful quantitative measures of the data's quality. However, in general the following metrics are a strong basis for such an assessment:

- Verification - (applies to computational data only) how accurately does the computation solve the underlying equations of the model for the quantities of interest?
- Validation - how much agreement is there between realizations of a model in experiment and computational, or, rarely, analytic, results?
- Uncertainty - what is the quantitative level of confidence in our predictions?
- Sensitivity - how sensitive are results to changes in inputs or upon assumed boundary conditions?

Similar, and perhaps more difficult, problems pertain to simulation data. While such data may be perfectly precise in a numerical sense, simulations typically rely on many parameters, assumptions, and/or approximations. In principle, if the above are specified, and the quantitative metrics meet user requirements, the data can be used with a high level of confidence. A similar approach to defining data quality was recently proposed within the context of the Nanotechnology Knowledge Infrastructure Signature Initiative within the National Nanotechnology Initiative [47].

An often posed question in the research community with regard to data associated with peer-reviewed journal articles is that of peer review of the data itself. Indeed, it

has been reported that approximately 50% of data being reviewed for submission to the American Mineralogist Crystal Structure Database contained errors [48]. The elements defined above represent the key criteria by which to judge the quality of the data. General pedigree and provenance information are typically conveyed in most research articles, though they may be provided in insufficient detail to reproduce the data. The remaining elements of validation, verification, uncertainty, and sensitivity are relatively loosely defined within materials science and engineering, and best practices have not generally been developed for each element, or, where developed, are not in widespread use. For further discussion on these topics, the reader is directed to [49] and [50].

Intellectual property and liability

There is quite a bit of complexity and even ambiguity with regard to the legal protections governing scientific data [51]. In general, scientific data are treated as facts and therefore not copyrightable under US law. However, the aggregation of the data into a single compilation or database may be copyrightable in the US. Additionally, and importantly, the codes, formats, metadata, data structures, or any 'added value' to the data could also be subject to copyright. Laws in other parts of the globe, particularly the European Union, add complexity to the situation. The EU's database directive, for example, protects the wholesale use of databases by other parties without permission.

There may be instances where the authors of a document may not want their data released immediately on publication of the supported manuscript. They may have very good, justifiable grounds to protect their data for some period following publication. One likely reason may be additional time required to file an invention disclosure related to the data. Another case may be that the authors are in the midst of writing another manuscript dependent on the same data. To account for these special cases, the associated publication should have allowance to grant the author an embargo period to protect the data for a short time after document publication. Typically, by granting an embargo, the author must post the supporting data to a repository prior to manuscript publication, but the data is not released to the public until the embargo period has expired. This is a standard practice in other technical disciplines, with limits of 12 months being typical and at the discretion of the editor.

Proprietary and export control restrictions may also affect the release of the metadata associated with the dataset and could warrant embargo or even permanent withholding of the entire metadata description. Take a researcher who has been provided a quantity of material by an industrial partner. The researcher may be free to report on a newly observed deformation phenomenon in the material with respect to its microstructure but may be restricted by the partner in providing proprietary details about how the material was processed. In this case, the metadata may not contain the full pedigree and provenance needed to reproduce the experimental results. Export control provides an analogous situation; the data may not be restricted, but the metadata needed to provide full pedigree and provenance may reveal export-controlled information [52]. Allowances for the withholding of metadata from publication must be in place, and these decisions to either accept the embargo or reject the dataset should be left to the reviewer and editor. It should be noted in publication policy that authors take full responsibility for review and release of proprietary and export-controlled information.

Given the discussion above regarding intellectual protection of data, policy regarding the requirements for licensure of data for reuse should be made clear. Of course, one must also consider where the data repository resides, so any policy may have somewhat limited scope. Creative Commons has developed a series of free copyright licenses for public use when sharing creative works [53]. For example, one option may be to require all deposited data to be covered by a CC-BY license, as defined by Creative Commons. A CC-BY grants free use of data by all parties, including for commercial use but does require attribution. However, other data repositories, such as the Dryad digital repository, have chosen to implement a Creative Commons Zero (CC0) license in order to remove any barriers to data reuse [13]. CC0 dedicates the work to the public domain and does not legally require attribution to the data source; instead Dryad relies on community norms for proper attribution of data. This option is particularly suitable in a case where a researcher uses data from hundreds or thousands of data sets in their work, making citation of all sources impractical. And, as noted at the beginning of this subsection, even the question of applicability of a copyright license to technical data is still open. Still, unanswered questions also linger regarding any liability issues with making data accessible. Again, consideration must be given to where the data reside (who is making it available) as to liability determination.

Archiving policy

A potential path forward is to establish a working group(s) comprised of members from the MSE community to craft a common data archiving policy. Such a policy should address the following:

1. A general definition of the data to be archived that is flexible enough to meet specific publication needs.
2. Criteria for suitable repositories.
3. Expectations or requirements to follow data or metadata standards.
4. Definition of standards for data citation and attribution.
5. Requirements and/or measures for data quality.
6. Clarity on intellectual property and liability issues.
7. Areas of opportunity for targeting pilot data archiving efforts (e.g., thermodynamic data).

Repositories

A complementary working group(s) from the MSE community should also be commissioned to develop a plan to provide supporting repositories for the MSE community. Some anticipated tasks and options include

1. Catalog and explore the suitability of and potential for existing materials repositories to host datasets associated with peer-reviewed journals (e.g., NIST CALPHAD database).
2. Explore the use of other established journal data repositories for their suitability for MSE data (e.g., www.datadryad.org).
3. Engage funding agencies for help in establishing specialized MSE data repositories.

4. Develop a time-phased strategy to provide well-structured materials repository architectures.
5. Consider business models that would sustain these repository services over the long term.

Conclusions

The era of open science is upon us, and the MSE community must generate a response that best suits the needs of not only the individual researcher but also the larger community including academia, industry, and government. It is becoming clearer with the advance of materials research that supporting data can no longer be kept invisible from a technical publication. This paper has outlined the key issues that will need to be considered as the community develops an approach to data archiving supporting publications. Charting the right course will take time, and much effort as it is quite complex. Fortunately, other technical disciplines have begun a path for us from which we can learn and capitalize. Some suggested community-based actions have been outlined that would help pave the way in setting a common approach to archiving of materials data.

Endnotes

[a]This is based on a query to the Astrophysics Data System, http://adsabs.harvard.edu/, for peer-reviewed papers mentioning either 'SSDS' or 'sloan' in the title or abstract of the paper. A query executed on 9 April 2014 resulted in 5,825 papers.

[b]CHW and JAW discussion with AAP, STM, AIP, ACS, Elsevier (2012).

[c]CHW discussion with A. Acharya, Google, Inc. (2012).

[d]'Certain commercial equipment, instruments, or materials (or suppliers, or software, ...) are identified in this paper to foster understanding. Such identification does not imply recommendation or endorsement by the US Government, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose'.

[e]CHW discussion with M. Whitlock, U. British Columbia (2012).

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
CHW structured the flow of the paper. CHW and JAW contributed a substantive portion of the manuscript, while RJH added valuable complementary perspectives from outside materials science and engineering throughout the subsections in the paper. All authors read and approved the final manuscript.

Acknowledgements
The authors wish to thank Clare Paul and Jeff Simmons for helpful discussions in preparing this manuscript.
This article has undergone an impartial review process. Charles H. Ward excused himself from any tasks associated with the processing of this article. The review process was handled entirely by other editors and all decisions regarding this publication have been made by them.

Author details
[1]Air Force Research Laboratory, Materials and Manufacturing Directorate, Wright-Patterson AFB, OH 45433, USA. [2]National Institute of Standards and Technology, Gaithersburg, MD 20899, USA. [3]Space Telescope Science Institute, Baltimore, MD 21218, USA.

Received: 20 May 2014 Accepted: 29 August 2014
Published online: 12 September 2014

References

1. Materials Research Society and The Minerals, Metals and Materials Society (2013) MRS-TMS 'big data' survey. JOM 65:1073, doi:10.1007/s11837-013-0724-y

2. Committee on Integrated Computational Engineering (2008) Integrated computational materials engineering: a transformational discipline for improved competitiveness and national security. The National Academies Press, Washington, DC
3. National Institutes of Health (2003) Final NIH statement on sharing research data. http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html. Accessed 13 May 2014
4. Department of Energy (2013) Policies on release of human genomic sequence data. http://web.ornl.gov/sci/techresources/Human_Genome/research/bermuda.shtml. Accessed 13 May 2014
5. National Science Foundation (2013) Grant proposal guide. http://www.nsf.gov/pubs/policydocs/pappguide/nsf13001/gpg_2.jsp#dmp. Accessed 13 May 2014
6. The White House (2011) The materials genome initiative. http://www.whitehouse.gov/mgi. Accessed 13 May 2014
7. Alliance for Permanent Access (2014) Opportunities for data exchange., http://www.alliancepermanentaccess.org/index.php/community/current-projects/ode/outputs/. Accessed 19 May 2014
8. The Royal Society (2012) Science as an open enterprise. https://royalsociety.org/~/media/policy/projects/sape/2012-06-20-saoe.pdf. Accessed 13 May 2014
9. RCUK Common Principles on Data Policy (2014) Research councils UK. http://www.rcuk.ac.uk/research/datapolicy/. Accessed 13 May 2014
10. Holdren JP (2013) Increasing access to the results of federally funded scientific research. http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf. Accessed 13 May 2014
11. The White House (2014) Project open data., http://project-open-data.github.io/. Accessed 13 May 2014
12. National Institutes of Health (2013) GenBank overview. http://www.ncbi.nlm.nih.gov/genbank/. Accessed 13 May 2014
13. Dryad digital repository (2014) Dryad. http://www.datadryad.org/. Accessed 13 May 2014
14. National Institute of Standards and Technology (2014) Thermodynamics Research Center. http://www.trc.nist.gov/. Accessed 13 May 2014
15. University of Leicester (2014) Peer REview for Publication & Accreditation of Research data in the Earth sciences [Online]. Available: http://www2.le.ac.uk/projects/preparde. Accessed 13 May 2014
16. The International Virtual Observatory Alliance (2014) Virtual observatory. http://ivoa.net/. Accessed 13 May 2014
17. Oxford Journals (2014) Database: The Journal of Biological Databases and Curation. http://www.oxfordjournals.org/our_journals/databa/about.html. Accessed 13 May 2014
18. Nature Publishing Group (2014) Scientific data. Macmillan. http://www.nature.com/scientificdata/. Accessed 13 May 2014
19. PLoS ONE editorial policies. PLoS One: http://www.plosone.org/static/editorial#sharing. Accessed 13 May 2014
20. Warren JA, Boisvert RF (2012) Building the materials innovation infrastructure: data and standards a Materials Genome Initiative Workshop. National Institute of Standards and Technology, Gaithersburg, MD, doi:10.6028/NIST.IR.7898
21. Whitlock MC, McPeek MA, Rausher MD, Rieseberg L, Moore AJ (2010) Data archiving. Am Nat 175:145–146, doi:10.1086/650340
22. Thomson Reuters (2014) Data citation index., http://thomsonreuters.com/data-citation-index/. Accessed 28 August 2014
23. (2012) Notes for authors. Acta Crystallogr C68:e3–e11, doi:10.1107/S0108270111047019
24. Koga N, Schick C, Vyazovkin S (2013) New procedures for articles reporting thermophysical properties. Thermochim Acta 555:iii, doi:10.1016/S0040-6031(13)00060-9
25. Miller M, Suter R, Lienert U, Beaudoin A, Fontes E, Almer E, Schuren J (2012) High-energy needs and capabilities to study multiscale phenomena in crystalline materials. Synchrotron Radiat News 25(6):18–26, doi:10.1080/08940886.2012.736834
26. Ember C, Hanisch R (2013) Sustaining domain repositories for digital data: a white paper. http://datacommunity.icpsr.umich.edu/sites/default/files/WhitePaper_ICPSR_SDRDD_121113.pdf. Accessed 13 May 2014
27. NIST Computational File Repository (2014) National Institute of Standards and Technology. http://nist.matdl.org/dspace/xmlui/handle/11115/52. Accessed 13 May 2014
28. Shade PA, Groeber MA, Schuren JC, Uchic MD (2013) Experimental measurement of surface strains and local lattice rotations combined with 3D microstructure reconstruction from deformed polycrystalline ensembles at the micro-scale. Integr Mater Manuf Innovation 2:5, doi:10.1186/2193-9772-2-5
29. Shade PA, Groeber MA, Schuren JC, Uchic MD (2013) 3D microstructure reconstruction of polycrystalline nickel micro-tension test. http://hdl.handle.net/11115/152 . Accessed 13 May 2014
30. Johns Hopkins University (2014) Research data management services at JHU. http://dmp.data.jhu.edu/. Accessed 15 May 2014
31. Labarchives (2014) LabArchives LLC. Carlsbad, CA, http://labarchives.com/. Accessed 13 May 2014
32. Figshare (2014) Figshare, London. http://figshare.com/. Accessed 13 May 2014
33. Computational Materials Data Network (2014) ASM International., http://www.cmdnetwork.org/content/cmdnetwork/. Accessed 13 May 2014
34. Cheung K, Hunter J, Drennan J (2009) MatSeek: an ontology-based federated search interface for materials scientists. IEEE Intell Syst 24:47–56, doi:10.1109/MIS.2009.13
35. Freiman S, Madsen L, Rumble J (2011) A perspective on materials databases. Am Ceram Soc Bull 90(2):28–32
36. Rumble J (1991) Standards for materials databases: ASTM Committee E49. In: Kaufman JG, Glatzman JS (eds) Computerization and networking of materials databases: Second Volume, ASTM STP 1106. American Society for Testing and Materials, Philadelphia, pp 73–83
37. ASTM International (1999) ASTM E1314-89(1999): Practice for structuring terminological records relating to computerized test reporting and materials design formats. ASTM International, West Conshohocken, PA
38. Rumble J (2014) E-Materials Data. ASTM International. Stand News. http://www.astm.org/standardization-news/perspective/ematerials-data-ma14.html. Accessed 15 May 2014
39. Austin T, Bullough C, Gagliardi D, Leal D, Loveday M (2013) Prenormative research into standard messaging formats for engineering materials data. Int J Dig Curation 8:5–13, doi:10.2218/ijdc.v8i1.245
40. Schmitz GJ, Prahl U (2014) ICMEg—the integrated computational materials engineering expert group—a new European coordination action. Integr Mater Manuf Innov 3:2, doi:10.1186/2193

41. Campbell CE, Kattner UR, Liu Z-K (2014) The development of phase-based property data using the CALPHAD method and infrastructure needs. Integr Mater Manuf Innov 3:12, doi:10.1186/2193-9772-3-12
42. Jackson MA, Groeber MA, Uchic MD, Rowenhorst DJ, De Graef M (2014) h5ebsd: an archival data format for electron back-scatter diffraction data sets. Integr Mater Manuf Innov 3:4, doi:10.1186/2193-9772-3-4
43. Uhlir PF (rapporteur) (2012) For attribution-developing data attribution and citation practices and standards. National Academies Press, Washington
44. Socha YM (2013) Out of sight, out of mind: the current state of practice, policy, and technology for the citation of data. Data Sci J 12f:13
45. Data Citation Synthesis Group (2014) Joint declaration of data citation principles. FORCE11. https://www.force11.org/datacitation. Accessed 14 May 2014
46. Data Cite and International Assocation of Scientific, Technical and Medical Publishers (2012) STM-DataCite joint statement. http://www.stm-assoc.org/2012_06_14_STM_DataCite_Joint_Statement.pdf. Accessed 14 May 2014
47. National Nanotechnology Initiative (2013) Nanotechnology knowledge infrastructure data readiness levels discussion draft. http://www.nano.gov/node/1015. Accessed 14 May 2014
48. Downs RT, Hall-Wallace M (2003) The American mineralogist crystal structure database. Am Mineral 88:247–250
49. Cowles B, Backman D, Dutton R (2012) Verification and validation of ICME methods and models for aerospace applications. Integr Mater Manuf Innov 1:2, doi:10.1186/2193-9772-1-2
50. Committee on Mathematical Foundations of Verification, Validation, and Uncertainty Quantification (2012) Assessing the reliability of complex models. The National Academies Press, Washington, DC
51. Madison M (2012) In: Uhlir PF (rapporteur) (ed) The future of scientific knowledge discovery in open networked environments. National Academies Press, Washington, pp 101–106
52. Ward CH (2013) Implications of integrated computational materials engineering with respect to export control. AFRL-RX-WP-TM-2013-0156. Defense Technical Information Center, Fort Belvoir, VA
53. Creative Commons (2014) Creative Commons. http://creativecommons.org/licenses/. Accessed 14 May 2014

doi:10.1186/s40192-014-0022-8
Cite this article as: Ward *et al.*: Making materials science and engineering data more valuable research products. *Integrating Materials and Manufacturing Innovation* 2014 **3**:22.

Submit your manuscript to a SpringerOpen journal and benefit from:

► Convenient online submission
► Rigorous peer review
► Immediate publication on acceptance
► Open access: articles freely available online
► High visibility within the field
► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com